

Counting Colours in Compressed Strings

Travis Gagie¹ and Juha Kärkkäinen²

Aalto University, Finland
travis.gagie@aalto.fi

University of Helsinki, Finland
juha.karkkainen@cs.helsinki.fi

Abstract. Suppose we are asked to preprocess a string $s[1..n]$ such that later, given a substring's endpoints, we can quickly count how many distinct characters it contains. In this paper we give a data structure for this problem that takes $nH_0(s) + \mathcal{O}(n) + o(nH_0(s))$ bits, where $H_0(s)$ is the 0th-order empirical entropy of s , and answers queries in $\mathcal{O}(\log^{1+\epsilon} n)$ time for any constant $\epsilon > 0$. We also show how our data structure can be made partially dynamic.

1 Introduction

Coloured range counting is a well-studied problem with applications in, e.g., computational geometry, database research and bioinformatics. For this general problem, we are asked to store a set of n coloured points in \mathbb{R}^d such that later, given an axis-aligned box, we can quickly count the number of distinct colours it contains. Most papers on this problem have focused on $d \geq 2$ dimensions (see, e.g., [5]); the upper bound for general static one-dimensional coloured range counting has not changed since 1995, when Bozanis, Kitsios, Makris and Tsakalidis [1] gave an $\mathcal{O}(n)$ -word data structure that answers queries in $\mathcal{O}(\log n)$ time. Recently, however, Gagie, Navarro and Puglisi [3] considered the special case in which the coloured points are the integers $1, \dots, n$. Storing these points is equivalent to storing a string $s[1..n]$ over an alphabet whose size σ is the number of distinct colours, such that later, given a substring's endpoints, we can quickly count how many distinct characters it contains.

Gagie et al. gave a data structure that takes $n \log \sigma + \mathcal{O}(n \log \log n)$ bits and answers queries in $\mathcal{O}(\log n)$ time. (In this paper \log means \log_2 .) Their solution is built on work by Muthukrishnan [8] about coloured range queries in strings. Muthukrishnan defined $C[1..n]$ to be the array in which each cell $C[q]$ stores the largest value $p < q$ such that $s[p] = s[q]$ (or 0 if no such p exists). He observed that $s[q]$ is the first occurrence of that distinct character in $s[i..j]$ if and only if $i \leq q \leq j$ and $C[q] < i$. Therefore, the number of distinct characters in $s[i..j]$ is the number of values in $C[i..j]$ strictly less than i . Gagie et al. noted that, if we store C in a wavelet tree [4], which takes $n \log n + o(n \log n)$ bits, then we can

count all such values in $\mathcal{O}(\log n)$ time; for details, see [6]. This is already a slight improvement over the bounds we achieve with Bozanis et al.'s data structure [1], but Gagie et al. showed it can be reduced to $n \log \sigma + \mathcal{O}(n \log \log n)$ by modifying the wavelet tree.

In Section 2 we describe a simple data structure that achieves essentially the same bound as Gagie et al.'s. In Section 3 we extend the ideas from Section 2 to build a data structure that takes $nH_0(s) + \mathcal{O}(n) + o(nH_0(s))$ bits, where $H_0(s) \leq \log \sigma$ is the 0th-order empirical entropy of s , and answers queries in $\mathcal{O}(\log^{1+\epsilon} n)$ time for any constant $\epsilon > 0$. This may be useful for applications such as tracking the unique visitors to a website, allowing us to count the unique visitors in any given interval. In Section 4 we show how our data structure can be made partially dynamic.

2 Simple Blocking

In this section we give a simple proof that, using two normal wavelet trees and a straightforward encoding of C , we need store only $(1 + o(1))(n \log \sigma + n \log \log n)$ bits to answer queries in $\mathcal{O}(\log n)$ time. Without loss of generality, assume $\sigma = o(n / \log n)$; otherwise, we achieve our desired bound by simply storing C in a single, normal wavelet tree. Our idea is to break s into blocks of length $\sigma \log n$ and encode the entry $C[q]$ differently depending on whether the previous occurrence $s[p]$ of the character $s[q]$ is contained in the same block. If $s[p]$ is contained in the same block as $s[q]$, then we write $C[q]$ as the $\lceil \log b \rceil$ -bit offset of p within the block; otherwise, we write it as the $\lceil \log n \rceil$ -bit binary representation of p . Notice that, for each block, there are at most σ entries of C encoded as $\lceil \log n \rceil$ -bit numbers.

We build a bitvector indicating how each entry of C is encoded, which takes $n + o(n)$ bits. We build one wavelet tree storing all the $\lceil \log b \rceil$ -bit encodings, which takes at most $n \log b + o(n \log b) = (1 + o(1))(n \log \sigma + n \log \log n)$ bits, and another storing all the $\lceil \log n \rceil$ -bit encodings, which takes at most $\sigma \lceil n/b \rceil \log n + o(\sigma \lceil n/b \rceil \log n) = n + o(n)$ bits. Notice that, if $s[q]$ is the first occurrence of that distinct character in $s[i..j]$ and $C[q]$ is encoded in $\lceil \log b \rceil$ bits, then $s[q]$ must be between $s[i]$ and the end of the block containing $s[i]$. We can count all such characters in $\mathcal{O}(\log b) = \mathcal{O}(\log \sigma + \log \log n)$ time using the bitvector and the first wavelet tree. We can count all the other first occurrences in $\mathcal{O}(\log n)$ time using the bitvector and the second wavelet tree.

Theorem 1. *Given a string $s[1..n]$, we can build a data structure that takes $(1 + o(1))(n \log \sigma + n \log \log n)$ bits such that later, given a substring's endpoints, in $\mathcal{O}(\log n)$ time we can count how many distinct characters it contains.*

Notice that, if $\sigma \geq \log n$, then Gagie et al.'s data structure is within a constant factor of being succinct and the data structure we just presented is within a factor of 2 of being succinct. If $\sigma < \log n$, then we can store s in a multiary wavelet tree [2], which takes $nH_0(s) + o(n)$ bits, and answer any query by enumerating the characters in the alphabet and, for each one, using two $\mathcal{O}(1)$ -time rank queries to see whether it occurs in the given substring.

Corollary 1. *Given a string $s[1..n]$, we can build a data structure that takes $2n \log \sigma + o(n \log \sigma)$ bits such that later, given a substring's endpoints, in $\mathcal{O}(\log n)$ time we can count how many distinct characters it contains.*

3 Multi-Size Blocking

In this section we extend our idea from the previous section so that, instead of encoding entries of C differently for only two block sizes — i.e., $\sigma \log n$ and n — we use many block sizes. In particular, we use $\mathcal{O}(\log \log n / \log(1 + \delta))$ different block sizes,

$$1, 2^{1+\delta}, 2^{\max((1+\delta)^2, 2)}, 2^{\max((1+\delta)^3, 3)}, 2^{\max((1+\delta)^4, 4)}, \dots, n,$$

where $\delta > 0$ is a value we will specify later. Also, for each block size b , we consider s to consist of about $2n/b$ evenly overlapping blocks,

$$s[1..b], s[b/2..3b/2], s[b+1..2b], s[3b/2+1..5b/2], \dots, s[n-b+1..n].$$

If $C[q] = p$ and the smallest block containing both $s[p]$ and $s[q]$ has size b , then we write $C[q]$ as the $\lceil \log b \rceil$ -bit offset of p within the leftmost of the (at most) two blocks of size b containing $s[q]$. Notice $\log b < (1 + \delta) \log(q - p) + 1$; calculation shows that the total size of all the offsets is at most $(1 + \delta)nH_0(s) + \mathcal{O}(n)$ bits.

Let t be a string indicating whether each entry of $C[q]$ is 0 and, if not, the block size used for it. We build a multiary wavelet tree [2] storing t . Since we can always encode a block size b using $\mathcal{O}(\log \log b)$ bits — even if δ is very small, thanks to the max in the definition of the block sizes — more calculation shows that $H_0(t) = \mathcal{O}(\log(H_0(s) + 1))$. It follows that, if $H_0(s)$ grows without bound as n goes to infinity, then the size of the tree is $o(nH_0(s))$ bits; otherwise, it is $\mathcal{O}(n)$ bits. Using the tree, in $\mathcal{O}(1)$ time we can count all the characters whose first appearance in s is in $s[i..j]$.

For each block size b , we build a wavelet tree storing all the $\lceil \log b \rceil$ -bit encodings. By the same calculation as for the offsets, these wavelet trees take a total of $(1 + \delta)nH_0(s) + \mathcal{O}(n) + o(nH_0(s))$ bits. Notice that, for any block size b , if $s[q]$ is the first occurrence of that distinct character in $s[i..j]$ and $C[q]$ is encoded in $\lceil \log b \rceil$ bits, then $s[q]$ must be between $s[i]$ and the end of the rightmost of the (at most) two blocks of size b containing $s[i]$. Using the multiary wavelet tree and the wavelet tree for block size b , in $\mathcal{O}(\log b)$ time we can count all such characters in the right halves of both the leftmost and the rightmost blocks of size b containing $s[i]$. Since the right half of the leftmost block is the left half of the rightmost block, the sum is the total number of such characters. It follows that we can count all the distinct characters in $s[i..j]$ in $\mathcal{O}(\log n \log \log n / \log(1 + \delta))$ time. Choosing $\delta = 1 / \log \log n$, for example, yields the following theorem:

Theorem 2. *Given a string $s[1..n]$, we can build a data structure that takes $nH_0(s) + \mathcal{O}(n) + o(nH_0(s))$ bits such that later, given a substring's endpoints, in $\mathcal{O}(\log n (\log \log n)^2)$ time we can count how many distinct characters it contains.*

A closer analysis shows that the time to count the distinct characters in $s[i..j]$ is $\mathcal{O}(\log(j-i+1) \log \log n \log \log(j-i+2))$. In a future version of this paper we will improve this bound to $\mathcal{O}(\log(j-i+1) + \min(\log(j-i+1), \log \log n)^2)$ without increasing our space bound. As far as we know, no other data structure for coloured range counting has a non-trivial upper bound depending only on the size of the range.

4 Partial Dynamism

Suppose $s[i_x]$ and $s[i_y]$ are the last occurrences of x and y strictly before $s[j]$, and $s[k_x]$ and $s[k_y]$ are their first occurrences strictly after $s[j]$. Then to change $s[j]$ from an x to a y , we need only reset $C[j] = i_y$, $C[k_x] = i_x$ and $C[k_y] = j$. To delete a character from s , we replace it with a special null character not in the alphabet (which we search for and exclude when performing queries). To append a character to s , we need only append an entry to C . Assume we have already found all the necessary positions using, e.g., a rank/select data structure for s (although, given some, we can find the others using our data structure from Section 3); in this paper we focus on how to update entries of C in our data structure's representation.

Mäkinen and Navarro [7] gave a dynamic data structure that stores a bitvector v of length n in $nH_0(v) + o(n)$ bits and supports rank, select, insert and delete in $\mathcal{O}(\log n)$ time. Using this dynamic bitvector data structure, they gave an efficient dynamic wavelet tree data structure. If we simply replace by standard dynamic wavelet trees the two static wavelet trees in our data structure from Theorem 1, then our space bound does not change and it takes $\mathcal{O}(\log^2 n)$ time both to count the number of distinct characters in a given substring and to update an entry of C .

If we simply replace with standard dynamic wavelet trees all the static wavelet trees (including the multiary wavelet tree) in our data structure from Theorem 2, then calculation shows we use $nH_0(s) + \mathcal{O}(n) + o(n(H_0(s) + \log \log \log n))$ bits and $\mathcal{O}((\log n \log \log n)^2)$ time both to count the number of distinct characters in a given substring and to update an entry of C . This space bound is $o(n \log \log \log n)$ bits larger than the space bound in Theorem 2 because t — the string indicating the block size used for each entry of C in Section 3 — is over an alphabet of size $\mathcal{O}(\log \log n / \log(1 + \delta))$. Therefore, whereas a multiary wavelet tree for t takes $nH_0(t) + o(n) = \mathcal{O}(n) + o(nH_0(s))$ bits, a standard wavelet tree for t (static or dynamic) takes $nH_0(t) + o(n \log \log \log n) = \mathcal{O}(n) + o(n(H_0(s) + \log \log \log n))$ bits. If we use a Huffman-shaped dynamic wavelet tree to store t , however, then it takes only $n(H_0(t) + 1) + o(n(H_0(t) + 1)) = \mathcal{O}(n) + o(nH_0(s))$ bits. We will give details in a future version of this paper.

Lemma 1. *We can make our data structure from Theorem 2 dynamic, without changing its space bound, such that it takes $\mathcal{O}((\log n \log \log n)^2)$ time both to count the number of distinct characters in a given substring and to update an entry of C .*

Theorem 3. *Suppose we have access to a dynamic rank/select data structure storing s such that queries, insertions and deletions all take $\mathcal{O}((\log n \log \log n)^2)$ time. Then we can build another data structure that takes $nH_0(s) + \mathcal{O}(n) + o(nH_0(s))$ bits such that in $\mathcal{O}((\log n \log \log n)^2)$ time we can replace, delete or append a character or, given a substring's endpoints, count how many distinct characters it contains.*

5 Acknowledgments

Many thanks to Veli Mäkinen, Giovanni Manzini, Gonzalo Navarro, Simon Puglisi and Jorma Tarhio, for helpful discussions.

References

1. P. Bozanis, N. Kitsios, C. Makris, and A. K. Tsakalidis. New upper bounds for generalized intersection searching problems. In *Proc. ICALP*, pages 464–474, 1995.
2. P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro. Compressed representations of sequences and full-text indexes. *ACM Transactions on Algorithms*, 3(2), 2007.
3. T. Gagie, G. Navarro, and S. J. Puglisi. Colored range queries and document retrieval. In *Proc. SPIRE*, pages 67–81, 2010.
4. R. Grossi, A. Gupta, and J. S. Vitter. High-order entropy-compressed text indexes. In *Proc. SODA*, pages 636–645, 2003.
5. H. Kaplan, N. Rubin, M. Sharir, and E. Verbin. Efficient colored orthogonal range counting. *SIAM Journal on Computing*, 38(3):982–1011, 2008.
6. V. Mäkinen and G. Navarro. Rank and select revisited and extended. *Theoretical Computer Science*, 387(3):332–347, 2007.
7. V. Mäkinen and G. Navarro. Dynamic entropy-compressed sequences and full-text indexes. *ACM Transactions on Algorithms*, 4(3), 2008.
8. S. Muthukrishnan. Efficient algorithms for document retrieval problems. In *Proc. SODA*, pages 657–666, 2002.